Application of semantic web technology to bio-informatics

Nilavamuthan Chandrasekaran and ShanmughavelPiramanayagam*
Computational Biology and Bioinformatics Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore- 641046, Tamil Nadu, India, nilavu.bioinfo@gmail.com & shanvel_99@yahoo.com

Abstract- The enormous amount of data resources are available on web, this makes integrative Bioinformatics research more important in life sciences research. Bioinformatics is about searching biological databases, comparing sequences, looking at protein structures, and solving biological and biomedical problems using a computer. Huge amount of online bioinformatics tools and data are available, so systems integration has become very important for further progress. Today, bioinformatics hosts heavily on the Web. But the Web is geared towards human interaction rather than automated processing. The approach of a Semantic Web facilitates this automation by annotating web content and by applying adequate reasoning languages. Semantic Web infrastructure utilizes the scalable Oracle Resource Description Framework (RDF) Data Model as the repository and Seamark Navigator for browsing and searching the data.
Keywords: Integrative Bioinformatics, Systems integration, Semantic Web, Biological databases Oracle RDF Data Model, Seamark Navigator.

Introduction

The Semantic Web is an evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content [1][2]. The Semantic Web is metadata based infrastructures for reasoning on the Web [3]. It extends the current Web without replacing it. Most of today's Web content is suitable for human consumption. Typical uses of the Web today involve humans seeking and consuming information, searching and getting in touch with other humans. The software tools to support these activities are not particularly well developed, only the search engines remains. The technology of these tools remains roughly the same, and Web content outgrows technological progress. The information retrieval is not very well supported. The major struggle is that, at present, the meaning of the Web content is not machine accessible [4], in the sense that computers cannot interpret words, sentences, and the relationships between them. To make possible the creation of the semantic Web the W3C (World Wide Web Consortium) has been actively working on the definition of open standards. These standards are important to define the information on the Web in a way that it can be used by computers not only for display purposes, but also for interoperability and integration between systems and applications resolving heterogeneity problems. Bioinformatics as a discipline has largely grown directly out of the molecular-biology laboratories where it was born. In general, each lab investigated a small region of biology and there are very few labs world-wide working on a single problem. Many of these labs have made their own data available for use on the Web. This data is often un- or semi-structured. Much of the data is composed of DNA or protein sequences, but this has generally been accompanied by large quantities of annotation of the sources of the sequences, literature citations and the possible function(s) of the molecules.

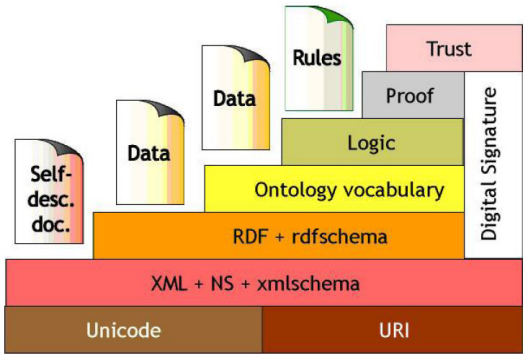

Fig. 1- The Semantic Web Stack from a 2000 presentation by Tim Berners-Lee *(http://www.w3.org/2000/Talks/1206-xml2k-tbl/)*

The ultimate aim of the Semantic Web is to add meaning to the current Web; it is to retrieve the meaning of data, the properties of objects, and the complex relationships existing between them by a series of formal rules, which would make information accessible to machines. Machine accessibility should be understood as representing information in such a way that it is possible to make queries based on the meaning of the data. The Semantic Web is an evolving collection of knowledge, built to allow anyone on the Internet to add what they know and find answers to their questions. Information on the Semantic web, rather than being in natural language text, is maintained in a structured form which is fairly easy for both computers and people to work with. Semantics is the study of relations between the system of signs and their meanings. Semantics have been defined in three forms [5].

1. Implicit semantics. "This type of semantics refers to the kind that is implicit in data and that is not represented explicitly in any machine processable syntax."
2. Formal semantics. "Semantics that are represented in some well-formed syntactic form (governed by syntax rules) is referred to as formal semantics."
3. Powerful (soft) semantics. "Usually, efforts related to formal semantics have involved limiting expressiveness to allow for acceptable computational characteristics. Since most KR mechanisms and the Relational Data Model are based on set theory, the ability to represent and utilize knowledge that is imprecise, uncertain, partially true, and approximate is lacking, at least in the base/standard models. Representing and utilizing these types of more powerful knowledge is, in our opinion, critical to the success of the Semantic Web. Soft computing has explored these types of powerful semantics. We deem these powerful (soft) semantics as distinguished, albeit not distinct from or orthogonal to formal and implicit semantics." The benefits promised by the Semantic Web include aggregation of heterogeneous data using explicit semantics, simplified annotation and sharing of findings, the expression of rich and Well-defined models for data aggregation and search, easier reuse of data in unanticipated ways, and the application of logic to infer additional insights [6]. To do work on web, we are searching in number of sites and downloading the information, this type of work is very expensive. The Semantic Web approach greatly simplifies this process. By using the web today, if you need data from 10 sites, we have to go to all 10 sites and cut and paste the data to get an integrated view, but by using Semantic Web Technology, it will assemble data out from the desktop into the network. With the Semantic Web, the network knows how to get and assemble the data. The Semantic Web browser will go and search from multiple sites to find the required information, retrieve this information, and display it in a single Semantic Web browser, this application of Semantic Web technology is a kin to a next-generation portal. Such capabilities make the Semantic Web very interesting to life science organizations. “The advent of the Semantic Web is providing the life sciences community with the standards and tools needed to build integrative informatics systems”.

The Semantic Web provides for defining and linking data to enable its more effective discovery, automation, integration, and reuse. It is to integrate disparate data both within and across functional areas, a promise that is generating strong interest from the life sciences community—for instance, the World Wide Web Consortium (W3C) has established its first Semantic Web interest group to focus on Health Care and Life Sciences (www.w3.org/2001/sw/hcls). The W3C standards recommendations that underlie the Semantic Web include RDF and OWL. Abstracting domain knowledge to an ontology layer avoids extensive reliance on custom procedural programming and the need to rewrite legacy code whenever a model, schema, or policy changes. RDF, OWL, and other Semantic Web technologies in development combine to provide a more effective mechanism to integrate data across drug discovery and development functions, promising a more supportive environment for earlier detection of safety-related issues. Eric Neumann has described early candidate areas for integration using Semantic Web technologies. Semantic Web standards have matured to the point where commercial software companies have implemented solutions. The latest release of the Oracle Database provides support as a repository for RDF-based information, including OWL. Oracle’s support generated much attention from a range of organizations that wanted to manage RDF data in a secure, scalable, and highly available environment. In addition, the Oracle Database now provides a single framework for managing and querying relational, XML, and RDF data. Cerebra have brought to market a scalable and robust solution for semantic information mediation and metadata interpretation. Cerebra Server provides highly optimized, decidable, and provable reasoning for OWL knowledge bases.

Objective

The objective is to create the core of a Bioinformatics Semantic Web populated by a number of sample data sources and applications representative of the use of the Web in Bioinformatics and to demonstrate novel, reasoning-based solutions dealing with the following problems:

1. Rules to formulate complex queries
2. Integration of Bioinformatics data
3. Adaptive portals for molecular biologists

Materials and methods

Biological data integration using Semantic Web technologies

The goal is to build a portal of gene-specific data to query and visualize, multiple information automatically mined from public sources. The features of the portal, called ‘‘Thea online’’ are similar to other gene portals like GeneCards [7], geneLynx [8], Source [9] or SymAtlas [10]. The technical solutions retained to implement the Web site should be totally transparent. So a centralized data warehouse was built in which all the data are aggregated in a central repository [11].

Gathering Data

The sources of information regarding human gene are collected. Available data are expressed in SWL, represented in tabular format or stored in tables in relational databases. Information expressed in SWL concerns protein centric data from UniProt [12], protein interactions data from IntAct [13] and the structure of Gene Ontology from GO [14]. These data are described in two different ontologies. UniProt and IntAct data are described in an ontology called core.owl. GO is a special case in the sense that it is not the definition of instances of an existing ontology, but it is ontology by itself in which GO terms are represented by classes. Data represented in tabular format concerns known and predicted protein protein interactions from STRING [15], molecular interaction and reaction networks from KEGG [16], gene functional annotations from GeneRIFs [17], GO annotations from GOA [18], literature information and various mapping files from NCBI [19]. Information from relational databases is extracted by performing SQL queries. This kind of information concerns Ensembl data [20] which are queried on a MySQL server at address ‘‘ensembldb.ensembl.org’’[11].

Data conversion

All the sources in future will be encoded in SWL, downloaded data will be imported directly in the data warehouse, but all the data that are not encoded in SWL needed to be converted. Tabular data were first converted in RDF with a simple procedure similar to the one used in YeastHub [21]. Each column which is to be converted in RDF was associated with a namespace that was used to construct the URIs identifying the values of the column. The relationship between the content of two columns was expressed in RDF by a triple having the content of the first column as subject, the content of the second column as object and a specified property. The conversions from tabular to RDF format were performed by dedicated Java or Python programs. The results obtained by SQL queries, which are composed of set of records, were processed the same way as data in tabular format [11].

Ontology of generated RDF descriptions

The vocabulary used in generated RDF descriptions is defined in a new ontology called Biowl. Classes (i.e.: Gene, Transcript, Translation) and properties (i.e.: interacts with, has score, annotated_with) are defined in this ontology using the namespace URI ‘‘http://www.unice.fr/bioinfo/biowl#’’ [11].

Data repository

Data collected from several sources which are associated with metadata and organized by ontology represent domain knowledge. As we chose a centralized data warehouse architecture, we need to store the set of collected and generated RDF/OWL specifications in a Knowledge Base. In order to be able to fully exploit this knowledge, we need to use a Knowledge Base System (KBS) [22] capable of storing and performing queries on a large set of RDF/OWL specifications (including the storing and querying of reified statements). It must include reasoning capabilities like type inference, transitivity and the handling of at least these two OWL constructs: ‘‘sameAs’’ and ‘‘inverseOf’’. In addition, it should be capable of storing and querying the provenance of information. At the beginning of the project, none of the existing KBS fulfilled these needs. The maximum amount of data handled by existing tools, their querying capabilities and the capabilities to handle contextual information were indeed below our needs (see the benchmark of several RDF stores performed in 2006 by Guo and colleagues [23]). For this reason, we developed and used a KBS specifically designed to answer our needs. Our KBS, called AllOnto is still in active development. It has been successfully used to store and query all the data available on the portal. At this time, it seems that Sesame version 2.0 (http://www.openrdf.org), released on December 20th 2007 has all the features allowing it to be equally used [11].

Querying using SPARQL

Triples stored in the KBS, can be queried with SPARQL queries.

Advantages of Semantic Web: [24]

Use of familiar, local terminology

Support for unanticipated modeling extensions

High degree of automation

High-fidelity integration and mapping with external systems and terminologies

Support for accurate answering of expressive queries

Conclusion & Future Work

The web has had an important effect on how science is now practiced: (1) research documents are rapidly distributed throughout a community for review and comment; (2) experimental data are easily shared with others, thus accelerating its analysis and interpretation; (3) internal databases containing the distillations of scientific research are publicly accessible and easily queried through user-friendly web pages; (4) scientific groups can share computational resources with each other through grid computing practices and (5) peer-reviewed journals offer online access allowing scientists to harvest large sets of relevant articles [6].

From this experiment, two main conclusions can be drawn:one which covers the technological issues, the other one which concerns more

sociological aspects. Thea-online is built on a data warehouse architecture [25] which means that data coming from distant sources are stored locally. It is an acceptable solution when the data are not too large and one can tolerate that information is not completely up-to-date with the version stored in source databases. However, the verbosity of SWL results in impressive quantities of data which are difficult to handle in a KBS. An import of the whole RDF serialization of UniProt has been successfully performed but improvements are still required in order to deal with huge data sets. From the technological point of view, the obstacles that must be overcome to fully benefit from the potential of Semantic Web are still important.

Acknowledgement
The authors gratefully acknowledge the facilities provided by DBT-BIF provided at Bharathiar University, Coimbatore, Tamil Nadu, India

References

[1] Tim Berners-Lee, James Hendler and Ora Lassila (2008). *Scientific American Magazine*.
[2] W3C (2008) *W3C Semantic Web Frequently Asked Questions,*03-13.
[3] Ivan Herman, W3C Head.*http://www.w3.org/2003/Talks/1112-BeijingSW-IH/*
[4] Loana Robu, Valentin Robu, Benoit Thirion. (2006) *J Med Libr Assoc*, 94(2): 198–205.
[5] Sheth, Ramakrishnan et al. (2005)
[6] Neumann E, Miller E., Wilbanks J:, (2004) *Biosilico* 228–236.
[7] Rebhan M., Chalifa-Caspi V., Prilusky J., Lancet D., (1998) *Bioinformatics* 14 656e664.
[8] Lenhard B., Hayes W.S., Wasserman W.W. (2001) *Genome Research,* (11), 2151-2157.
[9] Diehn M., Sherlock G., Binkley G., Jin H., Matese J.C., Hernandez-Boussard T.,Rees C.A., Cherry J.M., Botstein D., Brown P.O., Alizadeh A.A., (2003) *Nucleic Acids Research* 31219e223.
[10] Su A.I., Cooke M.P., Ching K.A., Hakak Y., Walker J.R., Wiltshire T., Orth A.P., Vega R.G., Sapinoso L.M.,Moqrich A., Patapoutian A., Hampton G.M., Schultz P.G., Hogenesch J.B. (2002) *Proceedings of the National Academy of Sciences* 99 4465 - 4470.
[11] Pasquier, *Institute of Developmental Biology and Cancer, CNRS e UMR 6543, University of Nice Sophia Antipolis, Parc Valrose, 06108 NICE Cedex 2, France*
[12] The UniProt Consortium (2007) *Nucleic Acids Research* 35 193 - 197.
[13] Kerrien S., Alam-Faruque Y., Aranda B., Bancarz I., Bridge A., Derow C., Dimmer E., Feuermann M., Friedrichsen A., Huntley R., Kohler C., Khadake J., Leroy C., Liban A., Lieftink C., MontecchiPalazzi L., Orchard S., Risse J., Robbe K., Roechert B., Thorneycroft D., Zhang Y., Apweiler R. (2007) *Nucleic Acids Research* 35 D561 - D565.
[14] Ashburner M., Ball C.A. , Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill L. Issel-Tarver D.P., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. (2000) *Nature Genetics* 25 25e29.
[15] von Mering C., Jensen L.J., Kuhn M., Chaffron S., Doerks T., Kruger B., Snel B., Bork P. (2007) *Nucleic Acids Research* 35 358e362.
[16] Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M., Hirakawa M., (2006) *Nucleic Acids Research* 34 D354eD357.
[17] Mitchell J.A., Aronson A.R., Mork J.G., Folk L.C., Humphrey S.M., Ward J.M. (2003) *American Medical Informatics Association, Annual Symposium Proceedings*, 460 - 464.
[18] Camon E., Magrane M., Barrell D., Lee V., Dimmer E., Maslen J., Binns D., Harte N., Lopez R., Apweiler R. (2004) *Nucleic Acids Research,*32, 262 - 266.
[19] Maglott D., Ostell J., Pruitt K.D., Tatusova T. (2007) *Nucleic Acids Research,* 35, 26-31.
[20] Birney E., Andrews T.D., Bevan P., Caccamo M., Chen Y., Clarke L., Coates G., Cuff J., Curwen V., Cutts T., Down T., Eyras E., Fernandez-Suarez X.M., Gane P., Gibbins B., Gilbert J., Hammond M., Hotz H.-R., Iyer V., Jekosch K., Kahari A., Kasprzyk A., Keefe D., Keenan S., Lehvaslaiho H., McVicker G., Melsopp C., Meidl P., Mongin E., Pettett R., Potter S., Proctor G., Rae M., Searle S., Slater G., Smedley D., Smith J., Spooner W., Stabenau A., Stalker J., Storey R., Ureta-Vidal A., . Woodwark K.C., Cameron G., Durbin R., Cox A., Hubbard T., Clamp M. (2004) *Genome Research,*14, 925-928.
[21] CheungKH., YipK Y., Smith A., deKnikker R., Masiar A., Gerstein M. (2005) *Bioinformatics,* 21, i85-i96.
[22] John M., Vinay C., Dimitris P., Adel S., Thodoros T. (1996) *The VLDB Journal* ,5, 238-263.

[23] Guo Y., Pan Z., Heflin J. (2004) *Proceedings of the Third International Semantic Web Conference*.
[24] Chimezie Ogbuji, Eugene Blackstone, Chris Pierce, (2007) *Cleveland Clinic*
[25] Surajit C., Umeshwar D. (1997) *ACM SIGMOD* ,26, 65-74.